Proc. Natl. Acad. Sci. USA
Vol. 91, pp. 6963–6966, July 1994
Biochemistry

A scheme for sequencing large DNA molecules by identifying local nuclear-induced effects

ITZHAK KELSON AND SHMUEL NUSSINOV

School of Physics and Astronomy, The Raymond and Beverly Sackler Faculty of Exact Sciences, Tel Aviv University, Tel Aviv 69978, Israel

Communicated by Yakir Aharonov, March 14, 1994

ABSTRACT **An experimental scheme for sequencing large DNA molecules is proposed where DNA strands are replicated, with all nucleotides of a given kind marked with radioactive ^{32}P. The marked strands are affixed to an appropriate substrate and are kept until most ^{32}P atoms decay. The local damage caused by the decay is expected to allow the identification of the sites occupied by that particular nucleotide, using atomic scale microscopy (scanning tunneling or atomic force microscopy). Quantitative aspects and methodological considerations associated with the proposed scheme are discussed.**

Large-scale DNA sequencing is one of the most intensive scientific efforts in biology today. In the present approach to sequencing (1, 2), one replicates DNA segments (with a common starting point) and selectively terminates them at sites occupied by a specific prechosen nucleotide type. The macroscopic ensemble of segments of different lengths thus obtained is electrophoretically separated into bands, and the positions of these bands are subsequently determined by standard radiography or fluorescence techniques.

An alternative, microscopic approach attempts to use the atomic scale tracking capabilities of scanning tunneling microscopy (STM) and atomic force microscopy (AFM) (3, 4) to identify *individual* nucleotides and to read directly the sequence of a *single* DNA molecule. To this end, considerable effort has gone into developing methods (5, 6) of stretching and adsorbing linear macromolecules onto surfaces and into finding the potentially promising parameters of STM or AFM application. At present, unfortunately, this approach is still impractical, primarily because of the lack of an unambiguous signature differentiating the four types of nucleotides from each other.

In this paper we propose an improved method of sequencing by atomic scale scanning, which utilizes much of the existing technical capabilities. The essential feature of this proposal is the way it makes use of radioactive nuclear decay. Marking with radioactive isotopes (e.g., with ^{32}P) is widely used to follow specimens and substances under study or to monitor them quantitatively. However, the nuclear decay itself can induce microscopic, local physical changes in the system in which it is embedded (7). It is on such effects that our proposition is based.

The Basic Idea

The basic idea implements a conceptually simple experimental scenario. The single-stranded DNA segment that is to be sequenced is replicated in such a way that *all* the nucleotides of one particular type (e.g., adenine nucleotides) are incorporated into the chain with radioactive ^{32}P. The marked molecules are then stretched and fixed onto a suitable substrate and are kept until most of the ^{32}P nuclei decay. The effect of this decay is twofold. First, the phosphorus atom transmutes into an atom of sulfur, which does not fit chemically into the original phosphorus site and bond configuration. Second, after the decay the nucleus (and the corresponding atom) recoils with up to 78 eV of kinetic energy (8), which is sufficient to physically remove it completely from its original site. Since the phosphates form the covalent links between consecutive nucleotides, this will cut the chain at this location (9)—namely, at the site of the ^{32}P-labeled adenine. We expect this induced break, as well as other local structural modifications, to be detectable by appropriately tuned STM (or AFM) scanning. By scanning along the chain, one can determine the length (or, equivalently, the number of nucleotides) separating consecutive adenine locations.

This entire procedure is repeated for chains with completely labeled cytosine, guanine, or thymine nucleotides in turn. The four sets of scans are then combined to yield the full DNA sequence.

Methodological Considerations

In evaluating the feasibility and the potential advantages of the proposed method, a number of methodological observations and considerations are appropriate.

(*i*) A number of auxiliary techniques that are essential for the implementation of the method are either already available or are the subject of intense development in conjunction with other methods. These include the fast replication of specific DNA chains (10–12), the stretching and fixation of such chains on suitably prepared substrates (13), and STM imaging. In particular, the ability to track by STM certain long polymer molecules and to identify their termination has been demonstrated (14).

(*ii*) The maximal total length, L, of a DNA chain that can be sequenced by a single application of any specific method is of primary importance to the overall sequencing project of very long DNA molecules. In such projects, one attempts to reconstruct the totality of the molecule from a multitude of much smaller segments. The complexity of this procedure, both in the preliminary biochemical sample preparation and in the subsequent combinatorial piecing together of the data, depends crucially on L. Increasing L results in an amplified reduction of this complexity. The currently applied "macroscopic" method is inherently limited by the decrease of resolution with the increase in the length of segments that are electrophoretically separated. This "band crowding" effect essentially imposes the limit (15)

$$L \leq 1000 \text{ in ordinary electrophoresis.}$$

What are the inherent limitations in the presently proposed method? Its application consists of two major generic operations. First, one has to measure repeatedly the distance between consecutive radioactive decay-induced local damage sites. This is a purely local measurement, which can be

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviations: STM, scanning tunneling microscopy; AFM, atomic force microscopy.

done accurately (with uncertainty smaller than one nucleotide unit) using the natural grid provided by the substrate. It is totally independent of the overall length of the chain. Second, one has to determine the absolute location along the chain of the series of distances thus obtained. This can again be done almost independently of L by combining low accuracy position determination of the scanned region (along the chain) with detailed local pattern matching.

The characteristic, practical limitations on L, differentiating this proposal from other, *microscopic* methods, are due to considerations related specifically to the ^{32}P usage and will be discussed below.

(*iii*) Carrier-free ^{32}P is available for the preparation of specific nucleotides, which are almost 100% marked with that isotope. We note that incomplete marking has simple quantitative consequences, which are discussed below. ^{32}P decays into ^{32}S with a characteristic lifetime τ_P of 20.6 days (or a half-life of 14.28 days). This decay effectively destroys the marked nucleotide as a possible building block of the replicated DNA chain. The replication process (which, for a single chain can be carried out within minutes) thus produces chains in which all nucleotides of the prechosen type are indeed marked with ^{32}P. In the time interval t_{fix} between the chain formation and its permanent fixing to the substrate, further decays will cause its breaking into shorter chains. For a chain of length L, of which one-quarter of the nucleotides are marked on the average, the survival probability in that time is given by $\exp[-(t_f/\tau_P)(L/4)]$. Thus, the length L depends primarily on the fixation time interval t_{fix} through

$$L \approx \frac{4\tau_P}{t_{fix}}.$$

For example, for $t_{fix} = 1$ h, $L \approx 2000$, while for $t_{fix} = 3$ min, which is conceivably realistic, $L \approx 40{,}000$.

(*iv*) Clearly, for the proposed method to succeed, we expect the radioactive decay (following fixation) to cause a well-localized effect but not to induce larger scale changes, such as desorption of neighboring nucleotides and possible modifications in their original linear ordering. Here we present the basic arguments in support of this expectation.

The recoil velocity v_0 of the nucleus ^{32}S resulting from the beta decay of ^{32}P is in the range of $0.5–2 \times 10^6$ cm/s (16). Note that v_0 is much smaller than typical electron velocities, which allows the electrons to adjust immediately to the instantaneous nuclear position and prevents excessive charging (beyond $Z = 1$ or 2) of the recoiling fragments. On the other hand, v_0 is larger than typical *nuclear* velocities and in particular exceeds the sound velocity along the DNA chain or in the underlying substrate. This is why the recoiling atom, whose kinetic energy is larger than its binding energy, is expected to actually break away from the chain, rather than generate a transient perturbation that propagates away and dissipates.

What effect would the recoiling ^{32}S atom have on neighboring elements (other atoms or whole nucleotides)? Let us consider generally such an element of mass M (in atomic mass units) at an initial distance r_0 from the decaying ^{32}P. Let us assume, for simplicity, that the ^{32}S recoils in a direction perpendicular to the chain and designate by h its position along this direction, at time t. The electrostatic force imparts to the mass M linear momentum, whose component in the perpendicular direction is given by

$$\delta P_\perp = \int F_\perp dt = Z_1 Z_2 e^2 \int (h/r^3) dt.$$

Z_1 and Z_2 are the effective charge states of the ^{32}S atom and of the element M, respectively. Substituting $dt = dh/v_0$ and $h\,dh = r\,dr$, we can readily integrate, getting

$$\delta P_\perp = \frac{Z_1 Z_2 e^2}{r_0 v_0}.$$

Substituting a value of 10^6 cm/s for v_0 and expressing r_0 in angstroms, we get for the recoil energy δE of the element of mass M

$$\delta E = \frac{\delta P_\perp^2}{2M} \approx Z_1^2 Z_2^2 \frac{m}{M} (r_0/\text{Å})^{-2} \text{ eV},$$

where m is the mass of the ^{32}S atom.

Let us consider first the nearby oxygen atom, which is double bonded to the ^{32}P atom, at a distance r_0 of ≈ 1 Å. Taking $Z_1 = Z_2 = 2$, reflecting the chemical bond nature, we get $\delta E \approx 20$ eV. Thus, this oxygen atom may conceivably be ejected along with the ^{32}S atom, forming possibly a simple SO compound.

On the other hand, for a neighboring nucleotide, viewed as a single element of mass $M \approx 200$ a distance of about 4 Å away, we get $\delta E \approx 0.01 Z_1^2 Z_2^2$ eV. (If any of the charges were to vanish, one would be left with much weaker Van-der-Waals type forces). For both effective charges Z_1 and Z_2 equaling unity, the resultant $\delta E = 0.01$ eV is smaller than the value of kT at room temperature. In any case, we expect the nucleotides to remain fixed to the substrate at their initial position. The only effect of the ^{32}P decay and the recoil of the resultant ^{32}S would be limited to its original site, with a possible local rearrangement of the frayed bonds of the chain.

This general conclusion is essentially a consequence of the removal of the ^{32}P atom as a chemical structure element from the chain and of the order of magnitude of energies involved. Although the detailed outcome of a particular decay event may depend on the initial direction of the recoiling atom (which we have taken, for simplicity, to be perpendicular to the chain direction), it is reasonable to expect the general validity of the qualitative, fraying effect.

The key question is, clearly, whether the effect of the decay can indeed be detected. Eventually, this question can only be positively answered by direct experimental demonstration. Such demonstration should compare identical sites before and after the ^{32}P decay has occurred. However, if it were possible to identify consistently and with atomic size resolution the location of the phosphorus atom and its neighboring contour, that, in itself, would provide strong *a priori* support for the detectability of the decay. Some experimental studies are definitely encouraging in that respect.

Driscoll *et al.* (17) have reported results of STM scans under ultrahigh vacuum conditions of a double-stranded DNA molecule ≈ 550 base pairs long. There is good quantitative agreement between the general geometric features of the observed image and parameters obtained by x-ray crystallography, including the helix pitch, the molecular width, the phosphate backbone width, and the axial nucleotide rise. The presumably A-type DNA image is compared with a model of the van der Waals surface derived from the crystallographic data, showing again both overall agreement of the qualitative features and a reasonable quantitative agreement of size and orientation. In particular, they compare interpolated experimental STM tip trajectories with corresponding atomic contours of that surface model, along different sections of the image. One clearly sees the regular agreement between the two, which holds specifically for the components of the phosphate backbone. An even more poignant example was obtained by Dunlap and Bustamante

(13), who imaged by STM *single-stranded* poly(dA) molecules on a highly oriented pyrolitic graphite surface. They demonstrate the ability to bind them to the surface with the deoxyadenylate molecules aligned in parallel to it and the phosphate backbones protruding upward. An image of a regular ensemble of two $(dA)_4$ strands is fitted by a detailed model of the structure, from which it is apparent that the phosphorus atoms are indeed identifiable with atomic resolution. Although this study is highly specialized (using, for example, non-self-complementary strands), it provides an example of the features of identifiability and reproducibility to which we referred above.

(*v*) One of the outstanding difficulties in microscopic scanning methods is the ambiguity in making positive identification of the DNA features. In fact, it was demonstrated (18) that scanning artifacts can be mistakenly identified as DNA strands. Our proposed method offers an intrinsic test of validity of the detection procedure and a direct quantitative check of its efficiency. This is based on the fact that the accumulated number of ^{32}P decays at a time t past the initial replication is simply proportional to $[1 - \exp(t/\tau_P)]$. Furthermore, by focusing the scanning onto a fixed area for a sufficiently long time, one should be able to view the actual appearance of the local damage caused by the decay, and, consequently, to search for the optimal conditions for its detection. As a simple example for the implementation of this point, consider a short, poly(A) single-stranded DNA molecule, in which all bases are marked with ^{32}P, which is affixed onto the scanned surface. Any modification of the observed atomic scale image of the chain is a clear indication of the effect of the ^{32}P decay and provides a direct means of its study.

(*vi*) Ideally, the application of the proposed method consists of fully scanning four independent copies of the same chain, each of which has been selectively marked with a different nucleotide type. In actual implementation the probability p of making a positive identification of the site of a specific nucleotide is smaller than unity for a number of reasons: incomplete marking of the original chain, partial decay of the marked sites, insufficient local damage, and microscopic detection efficiency. Furthermore, one should—as a general principle—employ strict criteria for acceptance of a positive signature, to eliminate the possibility of wrong, spurious site determinations.

The reconstruction of the complete sequence involves the need to fix, with high certainty, precise locations along the chain. If a given marked site is part of a sufficiently long congruous set of scanned nucleotides, then its location relative to other such segments can be readily established by conventional pattern recognition algorithms. This task is made particularly easy by the fact that the approximate position of any segment is intrinsically provided by the scanning procedure. Thus, it is enough to demand that any given nucleotide site is positively detected in a sufficiently large number of scanned molecules. Let N be the total number of scanned chains with a particular nucleotide type (say, adenine) marked. The probability that a specific adenine nucleotide site will *not be detected in any of them* is simply given by $(1 - p)^N$. A chain of total length L will have on the average $L/4$ sites of each particular type (adenine nucleotide in our example). Thus, to assure an even chance that a chain of total length L will have *no adenine nucleotide sites missed*, we require that probability to be smaller than $4/L$. Hence, N, the minimal required number of chains with a particular nucleotide marked, is given by

$$N \approx \frac{\log(4/L)}{\log(1-p)}.$$

For $L = 40{,}000$ and $p = 0.7$, for example, we get $N \approx 7$. Note that the requirement for a minimal number N of independent identification attempts of each site is valid even if they occur on smaller segments of the complete chain, provided their location can be reconstructed by suitable pattern matching. Such data segmentation might occur in practice, either because of inability to continuously scan the entire chain or by the actual breaking of the chain into smaller pieces.

It is interesting to point out a hypothetical variant of the method, which utilizes localized damage to the *substrate itself* induced by the impact of the recoiling atom. (In this case, one needs to remove the residual molecules from the surface prior to scanning). The corresponding intrinsic probability of detection, p, is expected to be much smaller for two reasons. First, the atom may recoil in an outward direction; second, its energy may not be enough to leave a permanent imprint on the surface. However, there are conceivable applications, where even low efficiency marking of the substrate might be useful, as in the determination of the geometric conformation to the surface of chains containing long poly(A) stretches.

(*vii*) Finally we note that, once the ability to fix the DNA molecules onto a suitable surface and to identify the nuclear decay induced damage with high probability had been developed, the entire scanning procedure can be carried out automatically.

To summarize, we have proposed a method for sequencing large DNA molecules that is based on the detection of local damage caused by radioactive decay at selectively marked sites. We have discussed the advantages of such a scheme (or, possibly, of variants based upon it) and presented the reasons for expecting it to work. Clearly, the proof of feasibility of the proposed method consists of a direct experimental demonstration of the detectability of that local damage and of establishing the accuracy of its position determination. We have, in fact, initiated an experimental program designed for this specific aim. However, this is a major effort, which involves testing a large number of possible experimental scenarios, including varying the nature of the substrate, its preconditioning for the stretching of DNA molecules, and the operational parameters of the scanning. We believe that expounding the idea at this stage is the most efficient way of advancing this required effort.

We would like to thank Prof. Moshe Schwartz for his substantive suggestions and critical comments.

1. Sanger, F., Nicklen, S. & Coulson, A. R. (1977) *Proc. Natl. Acad. Sci. USA* **74**, 5463–5467.
2. Maxam, A. M. & Gilbert, W. (1980) *Methods Enzymol.* **65**, 499–560.
3. Binnig, G., Rohrer, H., Gerber, C. & Wiebel, E. (1982) *Phys. Rev. Lett.* **49**, 57–60.
4. Engel, A. (1991) *Annu. Rev. Biophys. Chem.* **20**, 79–108.
5. Guckenberger, R., Hartman, T., Wiegräbe, W. & Baumeister, W. (1992) *Scanning Tunneling Microscopy II*, eds. Wiesendanger, R. & Güntherodt, H.-J. Springer Series in Surface Sciences (Springer, Berlin), Vol. 28, pp. 51–98.
6. Frommer, J. (1992) *Angew. Chem. Int. Ed. Engl.* **31**, 1298–1328.
7. Kelson, I. (1987) *J. Phys.* D **20**, 1049–1052.
8. Lederer, C. M. & Shirley, V. S., eds. (1977) *Table of Isotopes* (Wiley–Interscience, New York), 7th Ed.
9. Biggin, M. D., Gibson, T. J. & Hong, G. F. (1983) *Proc. Natl. Acad. Sci. USA* **80**, 3963–3965.
10. Mullis, K. B. & Faloona, F. A. (1987) *Methods Enzymol.* **155**, 335–350.
11. White, T. K., Arnheim, N. & Erlich, H. A. (1989) *Trends Genet.* **5**, 185–189.
12. Sambrook, J., Fritsch, E. F. & Maniatis, T. (1986) *Molecular Cloning: A Laboratory Manual* (Cold Spring Harbor Lab. Press, Plainview, NY).

13. Dunlap, D. D. & Bustamante, C. (1989) *Nature (London)* **342,** 204–206.
14. Arakawa, H., Umemura, K. & Ikai, A. (1992) *Nature (London)* **358,** 171–173.
15. Singer, M. & Berg, P. (1991) *Genes and Genomes* (Univ. Science Books, Mill Valley, CA), Chapt. 7.2.
16. Wu, C. S. & Moszkowski, S. A. (1966) *Beta Decay* (Wiley–Interscience, New York), p. 106–111.
17. Driscoll, R. J., Youngquist, M. G. & Baldeschwieler, J. D. (1990) *Nature (London)* **346,** 294–296.
18. Heckl, W. & Binnig, G. (1992) *Ultramicroscopy* **42–44,** 1073.